Toward Machine Understanding of Information Quality

Rong Tang

School of Information Science and Policy, University at Albany, SUNY. 113 Draper Hall, 135 Western Ave. Albany, NY 12222. Email: tangr@albany.edu

K. B. Ng

Graduate School of Library and Information Studies, Queens College, CUNY. New York, NY 11367
Email: kbng@qc.edu

Tomek Strzalkowski

ILS Institute, Univerity at Albany, SUNY. 1400 Washington Ave., Albany, NY 12222.
Email: tomek@albany.edu

Paul B. Kantor

School of Communication Information and Library Studies, Rutgers, the State University of New Jersey, 4 Huntington St., New Brunswick, NJ 08901. Email: kantor@scils.rutgers.edu

In this paper we report preliminary results of a study to develop, and subsequently to automate, new metrics for assessment of information quality in text documents, particularly in news. Through focus group studies, quality judgment experiments, and textual feature extraction and analysis, we were able to generate nine quality aspects and apply them in human assessments. Experts and students participated quality experiments, during which 1000 TREC documents were evaluated by participants from two sites -- Albany and Rutgers. Data showed good inter-judge agreement between judges from both sites. Principal component analysis revealed that the nine aspects form clusters of "content" and "presentation." Automatic quality prediction has been derived based on statistical analysis on the association between textual features and human quality judgments.

Introduction

In the information age, information quality is of significant importance to the user. Previous research has focused on identifying and itemizing key aspects of information quality in various contexts. Examples include the HON code developed for Medical Web sites (Health on the Net Foundation, 1997) and HITI's (Mitretek Systems, 1997) white paper on health information quality criteria. Meanwhile, the Kleinberg (1998) HITS algorithm and Google's PageRank technology (Brin & Page, 1998) were built on the rationale that hyperlinks are indications of Web page authoritativeness and quality. A number of studies addressed the issue of deriving automatic metrics to assess the quality of a given Web page (e.g., Price & Hersh, 1999; Amento, Terveen, & Hill, 2000; Zhu & Gauch, 2000). The problem remains, however, with regard to automatically capturing intrinsic quality features such as accuracy and objectivity, and to handling regular text-based news documents where the in-degree (number of citers) and out-degree (number of citees) measures are not necessarily present.

The information quality study reported here is a part of a large-scale multi-institutional project, named HITIQA (High-quality Interactive Question Answering). The overall goal of HITIQA is to perform advanced information retrieval and question answering through interactive dialogue and quality-based information fusion. Specifically, we are developing an extended model for classifying information by quality, in addition to, and as an extension of the traditional notion of relevance. The project involves Computer and Information science researchers from University at Albany and Rutgers University. Our intended clientele are intelligent analysts, and the documents that we targeted were news articles.

Related Research

The term "Quality" is defined by International Organization of Standards (1986) as "the totality of characteristics of an entity that bear on its ability to satisfy stated and implied need" (Standard 8402, 3.1). The concept of "Information Quality" has been addressed, at least theoretically, in numerous papers. Kahn and her coauthors (Kahn, et al., 2002) point out that the literature describes quality in four general ways: "as excellence, value, conformance to specifications, or meeting or exceeding consumer expectations" (p. 185). The value of information, as claimed by some scholars, lies in the quality of "fit for use" (Huang, et al., 1999; O'Brien, 1991). Other scholars declare that quality resides in meeting users' requirements (Kahn, Strong & Wang, 1998, Eppler & Wittig, 2000). Information quality, defined by O'Brien (1991), is "the degree to which information has content, form, and time characteristics which give it value to specific end users" (p. G-7). Upon evaluating seven conceptual frameworks, Eppler and Wittig state that there is a research need to examine interdependencies between different quality criteria and to develop actual tools based on an established framework. The authors further identified seven common elements of information quality, which include time dimension, accessibility, objectivity, relevancy, accuracy, consistency, and completeness.

There has been considerable research on classifying various quality aspects into broad categories. In a study by Wang and Strong (1996), data quality was classified into four categories: intrinsic, contextual, representational, and accessibility. Table 1 below shows the elements of each of the four quality dimensions. From a semiotics point of view, Helfert (2001) proposes that data quality characteristics be classified into the categories of Pragmatic (relevance, completeness, etc.), Semantic (accuracy, interpretability, reliability, etc.), and Syntax (syntactical correctness, consistency, accessibility, etc.). On the other hand, Naumann (2002) claims that quality criteria fall into four sets: content-related, technical, intellectual, and instantiation-related. *Content-related* criteria include accuracy, completeness, among others, and *Intellectual* criteria consist of believability, objectivity, reputation, etc. While availability and timeliness belong to the *Technical* dimension, representational conciseness and understandability are *Instantiation*-related.

TABLE 1. Information Quality Dimensions
(Source: Strong, Lee, Wang, 1997, p.39)

Category	Elements
Intrinsic IQ	Accuracy, Objectivity, Believability, Reputation
Accessibility IQ	Accessibility, Security
Contextual IQ	Relevancy, Value-added, Timeliness, Completeness, Amount of Information
Representational IQ	Interpretability, Ease of Understanding, Concise Representation, Consistent Representation

Empirical attempts to assess quality have been primarily focused on counting hyperlinks in a networked environment. Amento and his colleagues (2000) found that counts of simple in-degree (number of sites that link to the site) and out-degree (number of sites that the site links to) links worked well in identifying high quality Web sites. In another study, Price and Hersh (1999) devised automatic assessments of relevance, credibility, currency, bias, and content, in addition to a link-based measurement. The researchers used a set of specific words such as "mastercard" and "visa" to detect commercial bias. Content quality was measured by the proportion of text and ratio of hyperlinks to text. A third study by Zhu and Gauch (2000) used counts of in-degree links as a measure of some aspect of quality. Other measures included currency, availability, information-to-noise ratio, and authority. The authors computed cohesiveness of a Web page based on a formula that categorizes the page into the topics in a reference ontology using a vector space classifier. The results showed a significant improvement in search effectiveness when the six quality metrics were incorporated into a distributed search environment.

While hyperlinks and other web properties may help to derive a relatively sensible quality ranking of Web documents, these features do not apply to text-based news documents. Additionally, previous studies were only able to produce algorithmic measures for a limited number of quality aspects such as popularity. Measurements of many other important quality criteria such as accuracy, objectivity, depth, and conciseness have not been effectively automated. Our approach is to record human quality assessments of news articles and, based on this, conduct advanced statistical models of association between participants' quality scoring and occurrence and prevalence of certain textual features. In the end, we hope to induce automatic metrics of key quality dimensions for news documents.

Research Questions

Following are the three major research questions that we address:

1. What are the major quality aspects identified by news analysts as pertinent to their work?
2. What is the interdependency among the identified information quality criteria?
3. What is the association between textual features and the scoring of information quality aspects? In other words, how do we automatically predict the quality of any given document?

Methodology

Multiple research methods were used. Firstly, we conducted focus-group sessions to elicit key quality aspects. Secondly, we performed experts and students quality judgment experimental sessions. Thirdly, we identified a set of textual features, ran programs to generate counts of the features, and performed statistical analysis to establish the correlation between features and human quality ratings.

Focus Group Studies

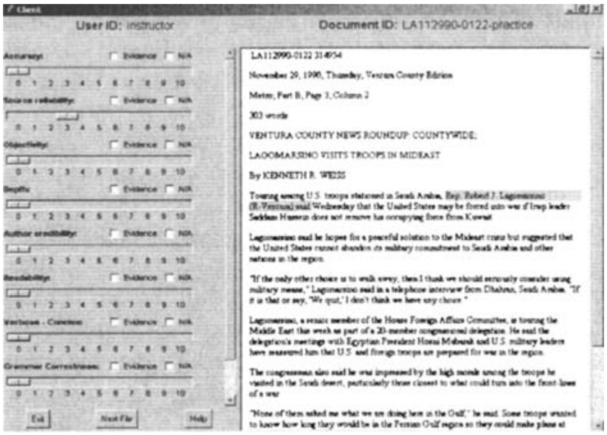
Two focus group sessions were conducted during March and April of 2002. The first session included seven participants who are news professionals with a majority of them from a local newspaper Albany Times Union. The second session included four professional news editors. Both sessions took 90 minutes, with the first more of a free-form discussion and the second more of a task-oriented meeting. Both focus group leaders followed the same general plan. The important criteria mentioned by participants of the first session include "source reliability," "objectivity," "completeness/context," "wording/nuances," "accuracy," "preciseness/veracity."

Participants of the second session first reported the top three quality aspects that they consider to be of great importance in their work. Each of the non repeated criteria was listed on the board, and they included "accuracy," "sources," "objectivity," "depth/detail/analysis," "author," "easy to read," "conciseness," "grammar/syntax," and "variety." There was much overlap between the criteria mentioned by the first and second session participants. In the second session, participants went through some reading samples, evaluated the document based on criteria discussed and highlighted segments of text as evidence for their assessment. In the end, participants worked together to rank the quality aspects in order of importance.

As a result of both sessions, nine quality aspects were generated. They are: accuracy, source reliability, objectivity, depth, author credibility, readability, verbose→conciseness, grammatical correctness, and one-sided→multi-views.

Quality Experiments

A computerized quality judgment system was created and the nine quality aspects were incorporated into the system, each with 10-point Likert scale. Figure 1 below is a screen shot of the system interface.



LA112990-0122 314954

November 29, 1990, Thursday, Ventura County Edition

Metro; Part B; Page 3; Column 2

303 words

VENTURA COUNTY NEWS ROUNDUP: COUNTYWIDE;

LAGOMARSINO VISITS TROOPS IN MIDEAST

By KENNETH R. WEISS

Touring among U.S. troops stationed in Saudi Arabia, Rep. Robert J. Lagomarsino (R-Ventura) said Wednesday that the United States may be forced into war if Iraqi leader Saddam Hussein does not remove his occupying force from Kuwait.

Lagomarsino said he hopes for a peaceful solution to the Mideast crisis but suggested that the United States cannot abandon its military commitment to Saudi Arabia and other nations in the region.

"If the only other choice is to walk away, then I think we should seriously consider using military means," Lagomarsino said in a telephone interview from Dhahran, Saudi Arabia. "If it is that or say, 'We quit,' I don't think we have any choice."

Lagomarsino, a senior member of the House Foreign Affairs Committee, is touring the Middle East this week as part of a 20-member congressional delegation. He said the delegation's meetings with Egyptian President Hosni Mubarak and U.S. military leaders have reassured him that U.S. and foreign troops are prepared for war in the region.

The congressman also said he was impressed by the high morale among the troops he visited in the Saudi desert, particularly those closest to what could turn into the front lines of a war.

"None of them asked me what we are doing here in the Gulf," he said. Some troops wanted to know how long they would be in the Persian Gulf region so they could make plans at

FIG.1 Quality Experiment Interface

One thousand medium-sized (100 words to 2500 words) news articles from the TREC collection (Voorhees, 2001) were used. They consisted of documents originally from LA Times, Wall Street Journal, Financial Time, and Associated Press with 25 relevant documents and 125 non relevant documents each from five Q&A topics of TREC.

We recruited two kinds of participants to perform judgments in sequence, experts and students. Expert sessions were completed first, and documents judged by experts were used for training student judges to perform quality assessment in accord with those of expert judges. The expert participants were experienced news professionals and researchers in the areas of Journalism, Communication, Political science, and Information Science. Expert sessions were held in May and June, and participants from both sites evaluated 100 documents. Each session took two hours. Participants were first informed about the purpose and procedure of the session. Each participant started with a practice document and then proceeded to the evaluation of ten documents. In the end, each of the 100 documents was rated twice, by two different experts, one at Albany, and one at Rutgers.

Ten documents judged by experts were selected and used as the training and testing material for student participants. The training and testing sessions were held in June and July. Student participants went through an initial orientation and read first five example documents, and the physical samples of experts' judgments. They then did their own ratings of the remaining five documents.

We compared the student evaluation with the expert's scoring and selected those participants whose judgments are more similar to that of experts by the following method. We used the sum of square of normalized differences between expert judgments and student judgment to test whether the student's judgment could have been drawn from the same (random) distribution as the expert judgment. Specifically, the following formula was used:

$$\frac{\left(\frac{x_{el}+x_{e2}}{2}-x_s\right)^2}{\left(x_{e2}-x_{el}\right)^2+1}$$

to measure the square of normalized difference between a student and an averaged expert, where x_{el} and x_{e2} are the scores of quality variable x assigned by the two experts respectively and x_s is the corresponding score assigned by the student. For the determination of the threshold, we approximated it by using the Monte Carlo method to identify the 95% point in observed distribution of this statistic, based on the observed mean and variance. If a student's sum of square of normalized difference exceeded the 95% point, we did not accept that student as a judge for further work. At Albany, 8 out of 45 tested students were disqualified whereas at Rutgers the ratio was slightly higher (4 out of 21).

Qualified student judges were then invited back for formal judgment sessions, which took place from June to August. Each session lasted two hours, during which participants evaluated ten documents. Each participant was required to do a minimum of two judgment sessions. In the end, participants as a whole evaluated 900 documents, and just as in the expert judgment process, each document was evaluated twice.

Textual Feature Extraction

With a collection of 1,000 judged documents, we began to establish a comprehensive set of textual features that seemed to be associated with the ratings of information quality. The textual features we considered include part of speech (POS), punctuation, vocabulary, length, use of certain key terms (See appendix A). We used existing POS taggers and information extraction tools developed for GATE (General Architecture for Text Engineering, see Cunningham et al., 2000) and AWB (Alembic Workbench, see Day et al, 1997), and some hand-built code to extract counts and features from the 1,000 documents.

We are currently in the process of data analysis, and will report some preliminary findings here. In our initial exploration, the statistical methods employed included principal component analysis (Reyment & Joreskog, 1993), discriminant analysis (Klecka, 1980) and logistic regression (Menard, 1995). Signal detection theory and the receiver operating characteristic curve (Egan, 1975) were also used to investigate the detection rate and false alarm rate for different predictive scores generated by the statistical analyses.

In the exploration stage, we concentrated on the prediction of the dichotomous value (i.e., high vs low) of the nine quality variables for a document. For example, in discriminant analysis, we sought a linear combination of the frequencies of various textual features (denoted by $\boldsymbol{v}_i$ in the following equation) as the basis for assigning cases into the two groups (high group vs low group):

$$\boldsymbol{D}_j = \beta_0 + \Sigma\beta_i \boldsymbol{v}_i$$

Where $\boldsymbol{D}_j$ is the discriminant score of the quality variable j (i.e., j ∈ { accuracy, source reliability, objectivity, depth, author credibility, readability, conciseness, grammatically correctness, multiple viewpoints }). Beta was chosen in such a way that the ratio of between-group sum of squares to the within-group sum of squares would be a maximum.

There are some inherent limitations associated with discriminant analysis. For example, it assumes a multivariate Gaussian distribution which can seldom be guaranteed for the array of variables based on textual characteristics. Therefore, we supplemented discriminant analysis with logistic regression:

$$\boldsymbol{p}_j / (1-\boldsymbol{p}_j) = \boldsymbol{e}^{\boldsymbol{y}}$$

where $\boldsymbol{p}_j$ is the probability for a document to have a high value of the quality variable $\boldsymbol{j}$, and $\boldsymbol{y}$ is a linear combination of the textual feature variables:

$$\boldsymbol{y} = \alpha_0 + \Sigma\alpha_i \boldsymbol{v}_i$$

The ratio between $\boldsymbol{p}_j$ and $\boldsymbol{1}-\boldsymbol{p}_j$ should be greater than one for those documents that have high value of $\boldsymbol{p}_j$ and less than one for those documents having low value of $\boldsymbol{p}_j$. Alpha was chosen in such a way that the number of correction predictions would be a maximum.

Results

Inter-judge Agreement

Figure 2 is the normality plot of the difference between scores assigned by Rutgers' judges and Albany's judges on the variable of "accuracy," with a mean almost equals to zero (with range from – 9 to + 9). As shown in the graph, the expected value of the difference is almost identical to the observed value. In other words, the difference between Albany and Rutgers judgments is normally distributed. The curves of the other eight quality variables are similar to the one below, all with diagonal curves and with means almost equal to 0.

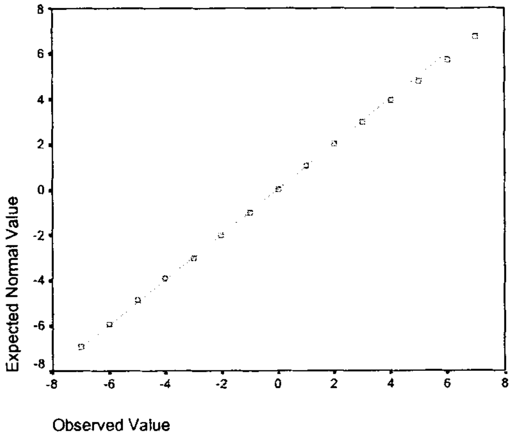


FIG. 2. Normality Plot of differences in quality judgments between Albany and Rutgers participants on the aspect of "Accuracy"

Dimensionality of Quality Criteria

Principal component analysis was conducted to detect the interdependency of the nine quality aspects used in judgment experiments. The first round of analysis yield the same two components from Albany data as from Rutgers data. The first cluster includes Accuracy, Source Reliability, Objectivity, Depth, Author Credibility, and Multi-views. The second component consists of Readability, verbose-conciseness, and Grammatical Correctness. Tables 2 and 3 provide details of the makeup of the two components.

TABLE 2. Principal Component Analysis of Albany Judgments, rotation converged in 5 Iterations. Two components explain 63% of variance

Pattern Matrix	Component 1	Component 2
Accuracy	.727	.200
Source Reliability	.702	.100
Objectivity	.735	.115
Depth	.852	-.181
Author Credibility	.713	.162
Readability	-1.933E-03	.854
Conciseness	-4.658E-02	.882
Grammatical Correctness	.221	.583
Multi-view	.757	-8.086E-02

TABLE 3. Principal Component Analysis of Rutgers Judgments, rotation converged in 5 Iterations. Two components explain 50% of variance

Pattern Matrix	Component 1	Component 2
Accuracy	.727	.200
Source Reliability	.702	.100
Objectivity	.735	.115
Depth	.852	-.181
Author Credibility	.713	.162
Readability	-1.9E-03	.854
Conciseness	-4.6E-02	.882
Grammatical Correctness	.221	.583
Multi-view	.757	-8.0E-02

During an attempt to further decompose the first content-oriented component, we found that the makeup of the sub-clusters of the component varies by judges' locations. While Albany's data showed that "Objectivity" belongs to a cluster with accuracy, author credibility and source reliability, Rutgers' data grouped "Objectivity" with depth and multi-views. Here we see some evidence from Albany data in support of Wang and Strong's grouping, i.e., Objectivity being intrinsic quality aspect together with accuracy and reputation. However, Rutgers data suggested that "objectivity" as a social construct, connected more, for these judges, with internal content elements such as depth and multi-views. Figures 3 and 4 illustrate the rotated component space of Albany and Rutgers.

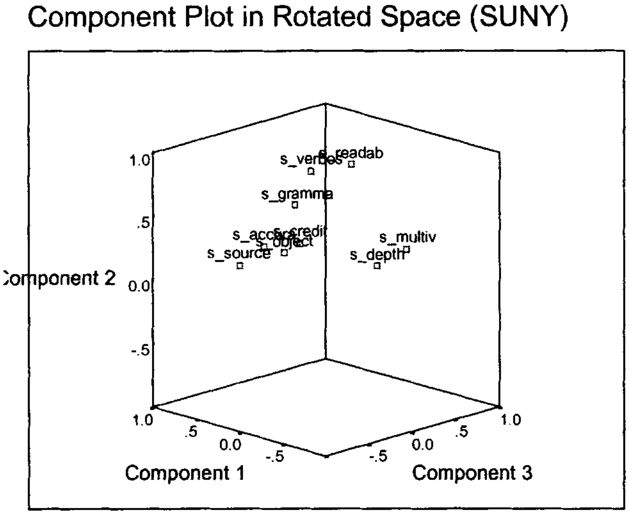


FIG. 3. Principal Component Analysis of Albany data

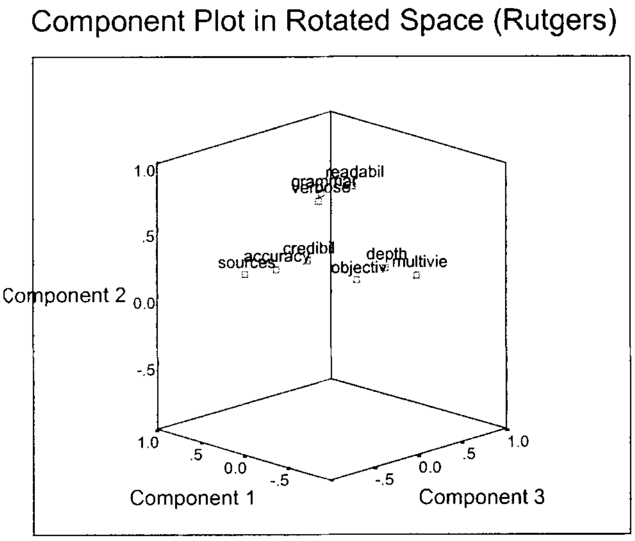


FIG. 4. Principal Component Analysis of Rutgers data

Quality Prediction

For all the judgment data, we recoded scores 1 to 5 as low and scores 6 to 10 as high. We wanted to see if it is possible to use textual features to predict whether a document was scored high or low in a particular quality.

We split the 1000 documents into two halves by random selection. In our training session the first half was used to estimate the parameters that would give best discriminant functions and logistic regression functions. In our testing session, we applied the functions to the other half to predict the nine quality criteria of the documents.

Table 4 is the summary of our preliminary findings. Logistic regression appears almost always better than discriminant analysis, but just by a very small amount:

TABLE 4. Performance of prediction (based on split- half training and testing) by two methods

	Discriminant Analysis Correct-Rate	**Logistic Regression Correct-Rate**
Accuracy	75.8%	75.9%
Source Reliability	67.8%	68.5%
Objectivity	70.6%	73.8%
Depth	77.4%	77.9%
Author Credibility	69.3%	71.7%
Readability	81.3%	83.0%
Conciseness	70.5%	70.9%
Grammatical Correctness	74.9%	75.1%
Multi-view	82.1%	82.2%

Altogether we had more than 150 textual features as independent variables, and we wanted to employ stepwise or backward method in our analyses to select the dominant predictive variables. In logistic regression, both stepwise and backward methods are very computational intensive. Since it turned out that the performance of using discriminant analysis and logistic regression are similar, we focused on stepwise discriminant analysis. The following is a summary of our preliminary finding:

TABLE 5. Quality Prediction by of Textual Features (from 5 to 17 variables selected by stepwise discriminant analysis)

Quality Variable	**Correct Prediction Rate**
Accuracy	68.5%
Source Reliability	56.9%
Objectivity	63.9%
Depth	66.9%
Author Credibility	55.1%
Readability	76.0%
Conciseness	63.0%
Grammatical Correctness	79.0%
Multi-view	69.6%

Mathematically speaking, the stepwise method does not give a better result, but it decreases the number of predictive variables dramatically. This is evident from the above table, when only 5 to 17 variables were used.

We can further illustrate this point by using "accuracy" as an example. When all the independent variables were applied in the discriminant analysis, the correct prediction rate is about 75%. However, when we decrease the number of predictive variables by stepwise discriminant analysis, the correct prediction rate decreased by only about 5%, and the number of predictive variables decreased to just six:

TABLE 6. Standardized Canonical Discriminant Function Coefficients of the quality variable: accuracy

Predictive Variable	**Coefficient**
Average length of paragraph in words	.675
Density of question mark	.335
Density of colon	.321
Density of the term "seem"	.228
Density of date	-.377
Superlative adjectives	-.455

Discussion and Conclusion

In answering our first research question, we found that there were nine quality aspects that were deemed important in news documents. Most of the nine aspects were mentioned in literature, however, it is important to single out the criteria that are pertinent to our research domain.

In answering the second research question, we found that the nine aspects appear to be grouped into two principal clusters, with one similar to Wang and Strong's intrinsic dimension and the other representational dimension. We named the first as "content" factor and the second "presentation" factor. Further decomposition of the content factor suggested that "objectivity" may be considered either as an internal content factor as in Rutgers' case or an external content factor as in Albany's case. This variation between the judges at two universities may reflect a difference in the paradigmatic stances of the two institutions. If news is taken to report a reality, then objectivity will be equated with something like "accuracy" or "source reliability." If the reality itself is seen as socially constructed, then the notion of objectivity becomes conflated with notions such as "fairness" or "balance," which in our case, are reflected through criteria "depth" and "multi-view."

In answering the third research question, we found that quality may be predicted based on textual features. Using statistical analysis alone, we can isolate 5 to 17 textual features to predict, with varying degrees of precision, different quality variables by linear combination of textual features. Most of the features we used were not based on semantics, which may be one reason why the prediction rates are relatively low. We believe that we could improve the performance if more semantic features are incorporated into the analysis. In our experiments, we have asked the judges to highlight and save the evidence supporting their quality judgments. These evidence texts may be a good source of more semantic variables. Besides using semantic features, many more statistical methods can be used for further studies. For example, in our analysis, we also found that backward logistic regression gave similar performance comparable to stepwise discriminant analysis, but based on a different subset of predictive variables (17 of them). There may be a way to combine the predictive powers of different methods by some form of data fusion.

In summary, we were able to identify important quality criteria relevant to intelligent analysts' work and we were also able to generate automatic quality metrics of news documents using users' quality judgments. Our next step is to apply our machine prediction method to produce measures for a new set of documents and have users to verify and modify machines' scoring. We hope that through this, we can collect new data to test our quality metrics and to further improve their performance.

ACKNOWLEDGEMENT

This paper is based on work supported by the Advanced Research and Development Activity (ARDA)'s Advanced Question Answering for Intelligence (AQUAINT) Program. The views expressed in this article are those of the authors, and do not necessarily represent the views of the sponsoring agency.

We gratefully acknowledge all the participants of quality experiments at both Albany and Rutgers. We also wish to thank Sharon Small, Robert Rittman, Ting Liu, Nobuyuki Shimizu, Peng Song, Ying Sun, Uzo Enyinna, and Tom Palen for their contributions to the project.

REFERENCES

Amendo, B., Terveen, L., & Hill, W. (2000). Does "authority" mean quality? Predicting expert quality ratings of Web documents. Proceedings of the Twenty-Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 296-303.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Proceedings of 7th International World Wide Web Conference (WWW7) / Computer Networks. 30(1-7), 107-117.

Cunningham, H., D. Maynard, K. Bontcheva, V. Tablan and Y. Wilks. (2000) Experience of using GATE for NLP R&D. In Workshop on Using Toolsets and Architectures To Build NLP Systems at COLING-2000, Luxembourg.

Day, D., Aberdeen, J., Hirschman, L., Kozierok, R. Robinson, P. & Vilain, M. (1997) Mixed-initiative development of language processing systems. In Fifth Conference on Applied Natural Language Processing, Association for Computational Linguistics. Retrieved on December 12, 2002 from : http://www.mitre.org/technology/alembic-workbench/ANLP97-bigger.html

Egan, J.P. (1975). Signal detection theory and ROC analysis. New York: Academic Press.

O'Brien, J. A. (1991). Introduction to Information Systems. (6th ed.). Boston, MA: Irwin.

Eppler, M. J., & Wittig, D. (2000). Conceptualizing information quality: A review of information quality frameworks from the last ten years. In Klein, B. D., Rossin, D. F. (ed.), Proceedings of the 2000 conference on information quality, Cambridge, MA: Massachusetts Institute of Technology, pp. 83-96.

Health on the Net Foundation. (1997). Code of Conduct for medical and health web sites, version 1.6. Retrieved November 9, 2002, from: http://www.hon.ch/HONcode/Conduct.html

Helfert, M. (2001). Managing and Measuring Data Quality in Data Warehousing. Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics. 55-65.

Huang, K. T., Lee, Y. W., & Wang, R. Y. (1999). Quality Information and Knowledge. Upper Saddle River, NJ: Prentice Hall.

Kahn, B. K., Strong, D. M., & Wang, R. Y. (2002). Information quality benchmarks: Product and service performance. Communications of the ACM, 45(4), 184-192.

Klecka, William, R. (1980) Discriminant analysis. Sage University Paper series on quantitative applications in the social sciences, series no. 07-019. Thousand Oaks, CA: Sage.

Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. Proceedings of 9th ACM-SIAM Symposium on Discrete Algorithms, 668-677.

Naumann, F. (2002). Quality-driven Query Answering for Integrated Information Systems. Berlin: Springer-Verlag. Retrieved on November 9, 2002 from: http://link.springer.de/link/service/series/0558/tocs/t2261.htm

Menard, Scott (1995) Applied logistic regression analysis. Sage University Paper series on quantitative applications in the social sciences, series no. 07-106. Thousand Oaks, CA: Sage.

Mitretek Systems. (1997). Criteria for assessing the quality of health information on the Internet. Retrieved on November 9, 2002, from: http://hitiweb.mitretek.org/docs/criteria.html

Price, S. L., & Hersh, W. R. (1999). Filtering Web pages for quality indicators: An empirical approach to finding high quality consumer health information on the World Wide Web. Proceedings of the AMIA 1999 Annual Symposium. 911-915.

Reyment, R. & Joreskog, K.G. (1993) Applied factor analysis in the natural science. Cambridge University Press.

Strong, D., Lee, Y., & Wang, R. Y. (1997). 10 potholes in the road to information quality. IEEE Computer, 30(8), 38-46.

Voorhees, E. (2001). Overview of TREC 2001. In E. Voorhees (ed.) NIST Special Publication 500-250: The Tenth Text REtrieval Conference, pp. 1 – 15. Washington, D.C.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of Management Information Systems, 12(4), 5-34.

Zhu, X., & Gauch, S. (2000). Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web. Proceedings of the Twenty-Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 288-295.

Appendix A. Text Features Considered

Punctuation	Number of periods, question marks, exclamation marks, commas, semicolons, colons, dash, ellipsis, parentheses, brackets, quotation marks, forward slides, apostrophes, hyphens
Symbol	Number of dollar signs, percent signs, plus signs, > marks, & marks
Length	Average length of words in characters, sentence in words, paragraph in words. Length of title, subtitle, leading paragraph, and document
Upper Case	Number of all upper case words, number of words with first character upper case
Quotation	Average quotation length
Key Terms	Number of word "say", "seem", and "expert"
Unique words	Number of unique words, number of unique words excluding stop words
POS	Number of token, proper noun, personal pronoun, possessive pronoun, determiner, preposition, verb in base form, verb in past tense, verb in present participle, verb in past participle, verb in present tense, verb in ing form
Entities	Number of person, location, organization, and date